KJMS 2025 VOL 7 (2): 207-216

KDU Journal of Multidisciplinary Studies (KJMS)
Volume 7. Issue 2 Nov 2025
DOI: https://doi.org/10.4038/kjms.v7i2.218

EXTENDING UNIVERSAL DEPENDENCIES TO TAMIL POETICS: MULTIWORD TOKENISATION AND ELLIPSIS IN THE THIRUKKURAḶ TREEBANK

Dilukshana, C[1] and Sarveswaran, K[2]

Department of Linguistics & English, Faculty of Arts, University of Jaffna, Sri Lanka[1]
Department of Computer Science, University of Jaffna, Sri Lanka[2]

ABSTRACT

Treebanks are critical resources in Natural Language Processing (NLP), supporting parser development, linguistic research, and the evaluation of large language models. While Tamil has seen progress in Universal Dependencies (UD) treebanking, existing corpora have been restricted to prose texts, leaving its vast poetic tradition underrepresented. This paper presents the first effort to be made to construct a syntactic treebank for Tamil poetry, specifically focussing on the ThirukkuRaḷ, which is composed in kuRaḷ veṇpā form. A central challenge in this work is posed by the treatment of multiword tokens (MWTs) and elliptical constructions, both of which are observed to occur frequently in Tamil verse due to its agglutinative morphology and metrical constraints. An annotation strategy is proposed within the Enhanced UD (EUD) framework to systematically address five major types of ellipsis—casal, verbal, adjectival, comparative/simile, and cumulative—alongside complex MWT patterns. These annotations not only enhance the representation of Tamil poetic syntax but also broaden the applicability of UD guidelines to underrepresented genres. The contribution is shown to underscore the linguistic and computational importance of capturing the structural specificities of Tamil poetry, while establishing a foundation for future cross-linguistic and literary treebanking efforts.

KEYWORDS: *Elliptical Compounds, Multiword Tokenization, Poetic Treebank, Universal Dependencies, Tamil Language*

Corresponding Author: Sarveswaran, K. email: sarves@univ.jfn.ac.lk

https://orcid.org/0000-0003-1579-0597

1. INTRODUCTION

A treebank is a linguistically annotated corpus in which each sentence is represented with its syntactic structure, typically in the form of trees. Treebanks have long been recognised as fundamental resources for natural language processing (NLP), serving as the backbone for parser development, syntactic evaluation, and theoretical linguistic inquiry (Marcus, Santorini and Marcinkiewicz, 1993). Beyond their role in grammar engineering, treebanks have also become indispensable in multilingual NLP and, more recently, in the evaluation of large language models (LLMs), where cross-linguistic syntactic benchmarks are required. For morphologically complex languages such as Tamil (Sarveswaran et al., 2021), treebanks are especially critical; yet Tamil remains under-resourced with respect to annotated corpora, part-of-speech (POS) taggers, and syntactically annotated treebanks (Bhattacharyya et al., 2019). This scarcity is observed to hinder both computational research and theoretical work on one of the world's oldest living classical languages (Hart, 2010).

Universal Dependencies (UD) (Nivre, 2005; de Marneffe et al., 2021) has been increasingly used in recent years as the framework for creating treebanks. More than 150 languages have adopted the framework for treebank development. Furthermore, the framework has been extended to encode additional linguistic information that is useful for training computational models and conducting linguistic analyses. Despite such advances, most existing UD treebanks remain restricted to prose texts, leaving poetic genres largely unrepresented. A few efforts have been made to annotate poetic corpora in languages such as Czech, English, and Finnish, but no systematic work has been undertaken for poetry in South Asian, or Dravidian languages.

This absence is particularly striking given the central cultural and linguistic value of Tamil poetry, with canonical works such as the *ThirukkuRaḷ* forming a cornerstone of Tamil literary heritage. The challenges of treebanking Tamil poetry are considerable: agglutinative structures, cliticisation, and metrical constraints often give rise to fused orthographic units that diverge from their syntactic decomposition. These units are treated as multiword tokens (MWTs) in Universal Dependencies terminology. Without systematic strategies for representing such phenomena, the computational processing of Tamil poetic text and its integration into UD work remain constrained.

Research Gap and Contribution

Multiword tokens are ubiquitous in Tamil, especially in poetic texts such as the *ThirukkuRaḷ*, yet no systematic syntactic annotation strategy has been developed for handling them. This gap not only impedes the computational analysis of Tamil poetry but also restricts the development of resources that capture the structural richness of Dravidian poetics. In this paper, this gap is addressed through the proposal of an annotation strategy for five major types of multiword tokens encountered in the *ThirukkuRaḷ,* designed within the UD framework. By doing so, the first syntactically annotated resource for Tamil poetry is introduced, and the relevant nuances are represented using the Enhanced UD specification.

1.1. **Background**

This section provides a brief overview of the Tamil language, its poetic structures — specifically the *kuRaḷ* poetic structure — and the *ThirukkuRaḷ* text.

Tamil

Tamil is a classical language spoken by more than 80 million people worldwide. It has a continuous literary tradition spanning several millennia and has evolved over time in both poetic and prose structures. Tamil is also a diglossic language, with significant differences between its spoken and written forms, and it exhibits a wide range of regional dialects (Schiffman, 1999). Its long-standing literary heritage is complemented by a parallel body of grammatical texts, developed across centuries, which document and analyse various linguistic aspects—including orthography, phonology, etymology, morphology, syntax, semantics, prosody, and contextual meaning (*Ilampūraṇanār et al.*, 2021).

The earliest grammar of Tamil, *Tolkāppiyam*, classifies literature into seven forms: *pāṭu* (poem), *urai* (prose), *nūl* (systematic treatise), *vaymoli* (oral instruction), *pisi* (riddles), *angatham* (satire), and *muthusol* (proverbs), all of which are attested in classical Tamil texts. Beyond stylistic variation, these forms also exhibit

substantial structural differences—particularly between poetry and other literary genres (Ilakkuvanar, 1963).

The focus of this work is poetry. A wide range of poetic types exists in Tamil, and each varies significantly in its structural and syntactic characteristics. In this paper, a Treebank is outlined for a specific poetic form known as *kuRaḷ veṇpā*.

kuRaḷ veṇpā

The genre known as *veṇpā* is a distinctive form of poetry that is governed by specific metrical rules. The *kuRaḷ veṇpā* in particular is composed of two lines. The basic unit of rhythm in *veṇpā* is called *asai*, which refers to a syllable or metrical foot. An *asai* may consist of one or two short or long letters (Al Asmath, 2021; Sri, 2023).

Each *asai* is composed of two major components: *ceer* and *thalai. Ceer* refers to the metrical foot, which is formed either by tokenising a single word or by concatenating multiple words according to the metrical structure of the poetic form (Madhavan et al., 2012). For instance, *kuRaḷ veṇpā* contains seven *ceer* (metrical feet) per line. Typically, words are grouped into feet, although closely connected words (such as those in apposition) may be grouped together as a single foot.

kuRaḷ veṇpā is written as a couplet consisting of four *ceer* in the first line and three *ceer* in the second line. Although additional rules must be followed for Tamil poetry, this paper does not examine the wider poetic system in further detail.

ThirukkuRaḷ

The *ThirukkuRaḷ*, one of the most celebrated works of Tamil literature, was composed by *Tiruvaḷḷuvar* and focuses on ethics and morality. The text is organised into three sections presenting aphoristic teachings on virtue *(aram)*, wealth *(poruḷ)*, and love *(inbam)*. It is widely recognised for its universality and secular orientation. Traditional accounts describe it as the final work of the Third Sangam, but linguistic analysis suggests a later composition date of 450–500 CE, after the Sangam period. The name *kuRaḷ* is traditionally assigned to this poet's great and only work, which comprises 133 chapters, each containing 10 couplets (Pope, 1980).

Tamil is an agglutinative language, where morphemes and words are concatenated into long, uninterrupted compounds. Although grammatical conventions exist for inserting spaces between words, they are not consistently observed—especially in poetic literature—in order to satisfy metrical requirements, thereby making multiword tokenisation a challenging task.

An additional complexity arises from the presence of *tokaiccol,* a type of compound in which certain morphemes or words are elided (Pope, 1980). Analysing these constructions is particularly difficult, as it requires identifying the specific type of elision involved.

2. METHODOLOGY

Since computational linguistics research is generally conducted within an applied research paradigm, this study has been designed following an applied research methodology. The methodological process follows a two-step approach: first, analysing classical Tamil poems, and then creating the treebank using the UD framework, as outlined below.

Classical Tamil poems were studied, with particular focus on the *veṇpā* poetic form, to identify their linguistic structures and relevant UD features.

- *ThirukkuRaḷ* texts were collected and analysed to examine their syntactic and morphological patterns.
- The structural components of *veṇpā* were identified.
- Part-of-speech categories were classified, using UPOS tags for the purpose of treebank construction.
- New UD dependency relations were identified, where necessary, to account for compound structures.
- POS tags, UD morphological features, and UD dependency relations were adapted to conform to the Universal Dependencies (UD) framework.

The following steps were followed in the construction of the Treebank using UD framework:

- Tokenisation: Each *kuRaḷ* was segmented into individual words (tokens).
- Multiword Tokenisation: Tokens were further segmented into smaller units where required.
- POS Annotation: Each token was assigned a UPOS tag.
- Morphological Feature Annotation: UD morphological features were annotated for all tokens.
- Dependency Relation Annotation: Syntactic relations between tokens were annotated using UD dependency relations.

3. LITERATURE SURVEY

Treebank annotation frameworks have been developed and refined over several decades, each contributing distinct insights into syntactic representation. The Penn Treebank was among the earliest resources to offer large-scale annotation of POS tags, phrase-structure syntax, predicate–argument relations, and disfluencies across multiple genres (Marcus et al., 1993). While influential, the Penn framework is limited by its language-specific design and its reliance on constituency-based annotation, which reduces its suitability for cross-linguistic comparison.

Building on this, the Prague Dependency Treebank (PDT) was introduced for Czech, offering a multilayer design that spans morphological, surface syntactic, and deep semantic (tectogrammatical) annotation (Hajič et al., 2020; Böhmová et al., 2003). PDT advanced the field by demonstrating that dependency grammars could model richer semantic phenomena. However, its complexity and language-specific layers restrict its scalability to other languages.

More recently, the Universal Dependencies (UD) project has emerged as the dominant cross-linguistic annotation framework, offering consistent, dependency-based, lexicalist annotation for more than 150 languages (May 2025 release) (Nivre, 2005; de Marneffe et al., 2021). UD's strength lies in its balance between typological comparability and language-specific flexibility. Nonetheless, critics note that UD's simplified, surface-oriented design sometimes struggles to capture deeper semantic or discourse-related features—particularly in morphologically rich and poetic languages. These limitations are especially relevant in the context of Tamil, where implicitness, ellipsis, and morphophonological fusion are prominent.

Universal Dependencies

Universal Dependencies (UD) is designed to provide a cross-linguistically consistent methodology for annotating syntactic structure (de Marneffe et al., 2021). Its primary goal is to establish a universal set of syntactic categories and relations that can be applied across diverse languages. Syntactic representation is encoded through dependency relations, where each word is linked to its syntactic head, forming a dependency tree.

Although UD offers a scalable, and widely adopted framework, several limitations have been noted in the literature. For example, UD's preference for surface syntax may be insufficient for analysing classical or poetic texts, where ellipsis, parallelism, and rhetorical structures are pervasive. Moreover, its treatment of compounds and multiword expressions can be ambiguous, particularly in languages with rich morphophonology such as Tamil.

Enhanced Universal Dependencies (EUD) addresses some of these limitations by extending dependency labels to represent semantic and implicit relations more clearly. EUD is particularly useful for handling ellipsis—an essential feature of Tamil classical poetry—and therefore offers a more expressive annotation layer for poetic corpora.

Within UD for English, one frequently used relation is compound, which describes the grammatical relation between two or more nouns that combine to form a single noun phrase. In such structures, the leftmost noun(s) serve as modifiers, while the rightmost noun functions as the head. For example, in a coffee store, coffee modifies store, and the UD annotation would be compound (store, coffee). This phenomenon is pervasive in English, which often employs NOUN+NOUN compounds to create new concepts (e.g., school bus, data analysis, student loan debt). Moreover, English allows such compounds to appear in multiple orthographic forms—open (osaka city), hyphenated (osaka-city), or closed (osakacity). Each variant is annotated consistently under UD

conventions, ensuring structural comparability despite surface variation, as shown in Figure 1.

However, this relation is less straightforward in Tamil, where compounding frequently interacts with morphophonological processes, *sandhi* rules, and metrical constraints. This mismatch highlights a key limitation of applying UD conventions directly to Tamil poetic structures without adaptation.

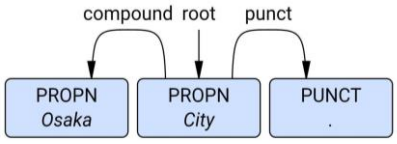

Figure-1: UD analysis of a NOUN+NOUN compound

Tamil UD Treebanks

Progress in Tamil UD resources has been limited but noteworthy. Three main Tamil treebanks have been introduced within the UD framework (Ramasamy, 2012; Krishnamurthy and Sarveswaran, 2021; Abirami et al., 2024; Sarveswaran, 2024).

The earliest resource, the Tamil Treebank (TTB), contains approximately 600 sentences drawn from news texts. While foundational, TTB includes automatically converted annotations, which reduces its reliability for high-quality syntactic research.

The Modern Written Tamil Treebank (MWTT) was subsequently developed with 534 manually annotated sentences selected from a modern Tamil grammar book. MWTT presents higher annotation consistency and serves as a gold-standard test corpus. However, its design as a small evaluation set limits its usefulness for training larger NLP models and its coverage remains restricted to prose.

A third resource, the Aalamaram Treebank (Abirami et al., 2024), extends Tamil syntactic annotation further, but it is currently unavailable online, limiting its accessibility and evaluability.

Critically, none of the existing Tamil UD treebanks include poetic texts, despite the central importance of poetry in the Tamil literary canon. The lack of treebanks for classical poetry represents a major gap, given that poetic language exhibits unique syntactic, metrical, and ellipsis-driven structures that differ profoundly from prose. This gap motivates the development of a Tamil poetry treebank, such as the *ThirukkuRaḷ* resource introduced in this study.

Multiword tokenisation

One of UD's key features is the treatment of multiword tokens (MWTs), defined as orthographic tokens that correspond to multiple syntactic words. UD guidelines require the systematic splitting of clitics, contractions, and related multiword phenomena (e.g. Spanish dámelo → da me lo) into their component syntactic words to ensure cross-linguistic consistency (Nivre et al., 2020). For agglutinative and clitic-rich languages such as Tamil, MWTs are particularly prevalent and must be handled with care (Stephen and Zeman, 2024). For instance, in the Tamil UD Treebank (TTB), 835 MWTs have been identified, with an average of 2.13 syntactic words per token, drawn from a total of 620 token types. This highlights the linguistic complexity of Tamil, where orthographic and syntactic boundaries frequently diverge.

Tamil is an agglutinative language, and syntactic words are often written together for stylistic or orthographic reasons. Therefore, multiword tokenisation has been recognised as an important component in UD treebanking for Tamil (Krishnamurthy and Sarveswaran, 2021; Ramasamy and Žabokrtský, 2011). Existing work, however, has primarily focused on prose and modern written varieties, leaving the interaction between metrical structure, orthography, and syntax in classical poetry largely unexplored. This gap is particularly significant for genres in which metrical constraints encourage the fusion or splitting of words in ways that are not directly addressed by current UD guidelines.

4. DISCUSSION

This section provides an account of the multiword tokenisation decisions adopted for the *kuRaḷ* Treebank and the elliptical constructions handled during annotation. These design choices were necessary to address the linguistic and metrical properties of

classical Tamil poetry, which frequently diverge from the assumptions underlying existing UD guidelines.

Multiword Tokenisation in the *kuRaḷ* Treebank

In classical Tamil poetry, the process of multiword tokenisation is especially crucial, as words are often joined into a single token or split into two tokens to satisfy metrical requirements. Consequently, before any syntactic analysis can be performed, both multiword tokenisation and, where necessary, concatenation must be applied to reconstruct the underlying syntactic structure accurately.

During the construction of the *kuRaḷ* Treebank, several recurrent multiword tokenisation patterns were identified, including Verb–Verb, Verb–Noun, Noun–Verb, Noun–Noun, Adjective–Noun, Numeral–Noun, and Noun/Verb–Clitic combinations. Multiword tokenisation in this context requires more than segmentation, as additional morphophonological transformations are often required due to the agglutinative and *sandhi*-rich nature of Tamil. Representative examples are provided below:

1) Verb –Verb concatenation
   eṇṇikkondattu
   eṇṇi -kondu -attu
   think.PART-take.PART- end.PAST.3SG
   "(He/it) ceased after considering (some- thing)."

2) Noun–Noun concatenation
   kāttalaritu
   kāttal -aritu
   act of preservation - rare
   "Reservation is rare."

3) Adjective – Noun Concatenation
   peruŋkaṭal
   perum -kaṭal
   "vast sea."

4) Numeral – Noun Concatenation
   aiṉtavittān
   aiṉthu + avittān
   "he who controlled the five (senses)"

5) Noun/ Verb - Clitic Concatenation, - ē
   mutaRē
   mutaRu + ē
   as the first cause / initiator

In metrical contexts, multiword tokenisation becomes even more critical, as orthographic forms frequently encode metrical rather than syntactic boundaries. When annotating syntactic relations—as illustrated in (3)—tokens must be split appropriately, taking into account assimilatory processes and *sandhi*. Conversely, some cases require that tokens be first concatenated and then re-segmented, because a single syntactic word has been divided across separate orthographic tokens, as demonstrated in (7).

There are also instances in which words cannot be further segmented, even though the same form may appear separately elsewhere in the corpus. The *ThirukkuRaḷ* provides many such examples, such as *vāḻvār* ('they will live') and *vāḻ vār*. Both forms occur in the text, and in cases such as *vāḻvār*, the token was retained as a single syntactic word when splitting would not have contributed additional syntactic information. This decision reflects a balance between maintaining internal consistency with UD tokenisation principles and preserving attested poetic usage, where orthographic variation serves metrical functions.

Overall, these tokenisation decisions extend current UD practice by permitting controlled pre-tokenisation concatenation, metrical-informed segmentation, and selective non-splitting. These choices highlight the limitations of existing Tamil UD guidelines for classical poetry and underscore the need for more explicit instructions for metrical languages.

Elliptical Constructions in the *kuRaḷ* Treebank

Tamil employs six major types of ellipsis, namely: casal ellipsis, ellipsis in similes or comparisons, verbal ellipsis, adjectival ellipsis, cumulative ellipsis, and ellipsis involving inexpressible or idiomatic terms (Pope, 1980). All five of the first categories are widespread in *kuRaḷ veṇpā*, where ellipsis is used strategically to align the text with metrical constraints. The sixth type—ellipsis involving inexpressible or idiomatic terms—carries deeper semantic content and is frequently excluded from syntactic analysis (Ilakkuvanar, 1963). Therefore, it is not addressed in this treebank.

Below, the first five elliptical constructions are outlined, together with their syntactic analyses and representation within the Enhanced Universal Dependencies (EUD) framework.

Casal ellipsis

Casal ellipsis (*vērrumait tokai* in Tamil) (Subrahmanya Sastri, 1934) refers to the elision of case markers. With the exception of the dative case (traditionally described as the fourth case in Tamil), most case markers may be elided in classical Tamil poetry.

A central challenge in annotating casal ellipsis lies not only in identifying the elided case, but also in determining the correct tokenisation, as compounds are often written together without spaces and without overt case marking. For instance, Figure 2 presents the compound noun *illanati* ('feet of the desireless one'), which is written as a single token in the poem. Simply assigning the entire token a nominal category would obscure its lexical and syntactic structure, thereby reducing its usefulness for downstream tasks such as machine translation or semantic interpretation. For this reason, the token must be segmented into the lemma forms *ilaan* and *ati*. In this example, the genitive case marking is elided, but the underlying relationship remains recoverable.

To address this, we propose to annotate such structures using the Enhanced Universal Dependencies (EUD) scheme, as illustrated in Figure 2. The enhanced annotation allows the implicit genitive relationship to be represented explicitly, thereby clarifying the syntactic role of each segment.

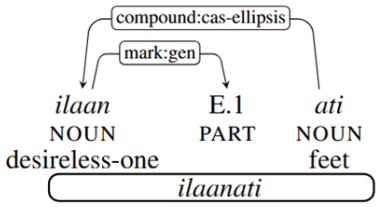

Figure 2: Example of Casal elision : *ilaanati*

As with the genitive example in Figure 2, other case markers may also be omitted, making it essential that they be annotated in the treebank to disambiguate their syntactic functions. Although human readers can typically infer the elided case relationships from context, this process is considerably more difficult for computational systems. Therefore, explicit annotation of casal ellipsis is necessary to ensure that parsing systems can interpret these constructions accurately and to support reliable downstream processing.

Verbal ellipsis

Verbal ellipsis — referred to as *vinaittokai* in Tamil (Subrahmanya Sastri, 1934) — occurs when a verb and a noun combine to form a compound in which the verb appears in its lemma or imperative form, without any explicit morphological marking such as tense or person–number–gender agreement. In such constructions, the verb functions adjectivally, modifying the noun that follows. As described in *Nannul* (verse 363) (Subrahmanya Sastri, 1934), when the tense or conjugational morphology of a verb is elided, the verb loses its inflection and acts as a descriptor that implicitly expresses all three tenses.

An illustration of this phenomenon is provided in Figure 3, where *virineer* is first tokenised and then annotated. In this compound, *viri* ('spread') + *neer* ('water') can denote three interpretations — water that has spread, water that is spreading, and water that will spread — as outlined in (6). Here, the future-tense marking is realised as an irregular or alternative form when the verb is compounded with a noun belonging to the *akirinai* category. Consequently, the usual participial suffix –a does not appear, illustrating how poetic compounding may suppress expected inflectional material.

These structures demonstrate how verbal ellipsis in classical Tamil poetry compresses semantic and syntactic information, relying on contextual and metrical cues for interpretation. Accurately recovering such elided features requires enhanced annotation, making the use of EUD particularly valuable for representing the implicit temporal and descriptive relationships encoded in *vinaittokai* compounds.

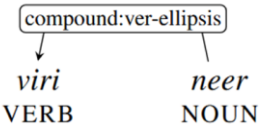

Figure 3: Example of Verbal-elision

(6) *viri* *neer*
spread. IMP water. NOM
viri-nt-a *neer*

spread- PAST -PARTICIPLE water. NOM
viri-kintr-a neer
spread- PRES -PARTICIPLE water. NOM
viri-yum neer
spread- FUT.PARTICIPLE water. NOM

A particular challenge in poetic texts is that the verb may appear at the end of one token, while the corresponding noun occurs in the subsequent token. When analysing such constructions, it is essential that the annotation indicates that the verb in the preceding token and the noun in the following token together form a single compound. For example, in *ThirukkuRaḷ* – 5, as shown in (7), *sēr* ('add') and *pukazh* ('praise') constitute a compound, despite being distributed across two separate orthographic tokens. In such instances, marking verbal ellipsis becomes especially important, as illustrated in 3, since the syntactic and semantic unity of the construction is not visible from the surface token boundaries alone.

(7) *irulsēr iruvinaiyum sēraa iraivan*
porulsēr pukazh purindhār māttu .

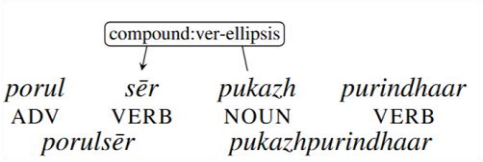

Figure 4: Verbal-elision with multiword tokenisation

Cumulative ellipsis

Cumulative ellipsis, referred to as *ummaittokai* in Tamil, arises from the elision of the clitic *–um*, which has eight different syntactic functions, one of which is to act as a conjunction marker (Subrahmanya Sastri, 1934). When *–um* functions as a conjunction, it is frequently elided in poetry to satisfy metrical constraints. Such elided constructions are classified as *ummaittokai* and require explicit annotation to preserve their syntactic interpretation.

In (8), two senses are listed without any overt conjunction marker. Furthermore, the names of the senses appear in their lemma form and are simply concatenated to adhere to the metrical structure of the poem. As a result, each item must first be tokenised through multiword tokenisation, after which the elided conjunction should be marked explicitly, as shown in Figure 5.

Marking the conjunction in such cases is essential, because the sequence may otherwise be misinterpreted. For instance, if the elided *–um* is not annotated, the construction could be mistakenly analysed as a case-elliptical compound rather than a coordinated structure. The explicit marking of the conjunction therefore prevents ambiguity and ensures that the syntactic relation between the coordinated nouns is made clear.

It should also be noted that Tamil contains another conjunction marker, *mattum*. However, in elliptical constructions within classical poetry, the *–um* clitic is typically the form that is elided, making it the primary focus of annotation in cumulative ellipsis.

(8) *venduthal vendamai*

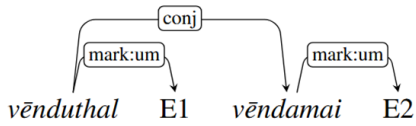

Figure 5: Example of - *um* elision or conjunction mark elision Given text should be written as *vēndutal-um vēndamai-yum*

Comparative/Simile ellipsis

Comparative ellipsis, known as *uvamait tokai* in Tamil, refers to constructions in which two or more words or phrases are juxtaposed, but the comparison marker—typically equivalent to 'as' or 'like'—is omitted (Subrahmanya Sastri, 1934). A similar phenomenon arises in simile ellipsis, where metaphorical comparisons are expressed without an overt comparative particle. Both forms are common in classical Tamil poetry, where metrical pressures frequently lead to the omission of such markers.

An example is provided in (9), where the compound *porivayil* must be interpreted as 'a door-like organ'. In this construction, the comparative marker is elided, and the two components are written together as a single token. For accurate treebank representation, the compound must first be tokenised, after which the implicit comparison is restored through annotation. This process enables the comparative or metaphorical relationship between the two elements to be captured

explicitly, as illustrated in (5), following the Enhanced Universal Dependencies framework.

By marking such ellipses, the syntactic and semantic relationship between the compared elements is preserved, preventing misinterpretation and ensuring that metaphorical or comparative structures in *kuRaḷ veṇpā* are represented faithfully within the treebank.

(9) *porivayil*
pori-vayil
Organ-door
'Doors like organ'

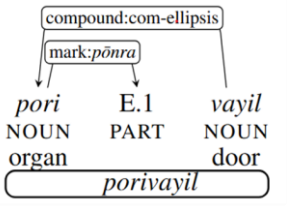

Figure 6: Example of Comparative or Simile ellipsis elision

Adjectival ellipsis

Adjectival ellipsis, referred to as *panputtokai* in Tamil, denotes a compound in which the first element expresses a quality—such as colour, shape, size, taste, or other inherent attributes—while the second element is a noun that embodies that quality. In this construction, the particle that would typically link the attribute to the noun—*'ākiya', 'āna',* or *'mai'*—is omitted, resulting in a surface form where one noun directly modifies another, functioning effectively as an adjective.

In such cases, the absence of the linking particle compresses the syntactic structure, creating a compound in which the attributive relationship must be inferred. For treebank annotation, this requires that the compound be tokenised appropriately and that the implicit adjectival connection be made explicit through Enhanced Universal Dependencies. By doing so, the underlying attributive function is preserved, ensuring that the syntactic and semantic relationship between the elements is represented faithfully, despite the absence of an overt marker.

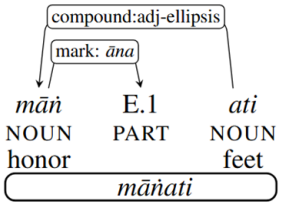

Figure 7: Example of adjectival elision : *mānati* = 'great/divine feet'

5. CONCLUSION

This article has presented the annotation of five elliptical linguistic constructions encountered during the development of the first poetic treebank for Tamil. These constructions have been annotated within the Universal Dependencies (UD) framework, using the Enhanced Universal Dependencies (EUD) specification, as part of the *ThirukkuRaḷ* Treebank project. The study has underscored the central role of multiword tokenisation in Tamil treebanking, particularly in poetic contexts where syntactic words are frequently fused to satisfy metrical constraints. Five major types of ellipsis—casal ellipsis, comparative and simile ellipsis, verbal ellipsis, adjectival ellipsis, and cumulative ellipsis—have been analysed, and new annotation labels (cas-ellipsis, ver-ellipsis, com-ellipsis, and adj-ellipsis) have been proposed to represent these constructions more accurately within the UD framework.

The novelty of this work lies in extending syntactic annotation beyond prose into the domain of classical Tamil poetry, a genre that presents distinctive challenges due to its agglutinative morphology, metrical structure, and extensive use of elided forms. While treebanking efforts for Tamil prose already exist, this study represents the first systematic attempt to annotate poetic texts, thereby addressing a significant gap in both Tamil NLP and cross-linguistic treebanking research. By capturing phenomena such as fused multiword tokens and elliptical constructions, the annotations developed here contribute not only to computational processing but also to a deeper understanding of the structural richness of Tamil poetics.

This work aims to advance the computational modelling of underrepresented languages and genres,

and to establish a foundation for the integration of poetic texts into UD resources. Future work will include scaling the annotation to cover all 1,330 couplets of the *ThirukkuRaḷ*, extending the framework to additional Tamil poetic forms, and conducting inter-annotator agreement studies to ensure consistency and reliability. Beyond Tamil, the annotation strategies introduced here have potential applications in comparative and cross-linguistic research, particularly for other agglutinative and clitic-rich languages with strong literary traditions. In the longer term, the resource developed through this project can support both linguistic inquiry and the culturally sensitive evaluation of large language models.

6. REFERENCES

Abirami, A.M., Leong, W.Q., Rengarajan, H., Anitha, D., Suganya, R., Singh, H., Sarveswaran, K., Tjhi, W.C. & Shah, R. (2024, May) 'Aalamaram: A large-scale linguistically annotated treebank for the Tamil language', *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pp. 73–83.

Al Asmath (2021) *Yāppiyaluṟai: Āyv uṭaṉāṉa oru vilakkam*, Vellampittiya Āciriyar edition (in Tamil).

Bhattacharyya, P., Murthy, H., Ranathunga, S. & Munasinghe, R. (2019) 'Indic language computing', *Communications of the ACM*, 62(11), pp. 70–75.

De Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021) 'Universal dependencies', *Computational Linguistics*, 47(2), pp. 255–308.

Hart, G.L. (2010) 'Credentials of Tamil as a classical language', in *Proceedings of the World Classical Tamil Conference*. Chennai.

Ilakkuvanar, S. (1963) *Tholkāppiyam in English: With Critical Studies*. "Kuṟaḷ.Neṟi" Publishing House.

Ilampūraṇanār, Naccinārkkiṉiyar, Cēṉāvaraiyar, Teyvachilaiyar, Kallāṭanār, Pērāciriyar & Rā. Ilaṅkumaraṇar (2021) *Tolkāppiya uraittokai* (vols. 1–19). Tamil edition (in Tamil). Chennai: Thamizhmann Pathippagam.

Krishnamurthy, P. & Sarveswaran, K. (2021, December) 'Towards building a modern written Tamil treebank', *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pp. 61–68.

Madhavan, K.V., Nagarajan, S. & Sridhar, R. (2012) 'Rule based classification of Tamil poems', *International Journal of Information and Education Technology*, 2(2), p. 156.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993) 'Building a large annotated corpus of English: The Penn Treebank', *Computational Linguistics*, 19(2), pp. 313–330.

Nivre, J. (2005) 'Dependency grammar and dependency parsing', *MSI Report*, 5133(1959), pp. 1–32.

Pope, G.U. (1980) *The Sacred Kurral of Tiruvalluvanayanar*. New Delhi: Asian Educational Services.

Pope, G.U. (2006) *Tamil Handbook: A Handbook of the Tamil Language*. New Delhi: Asian Educational Services.

Ramasamy, L. & Žabokrtský, Z. (2011, February) 'Tamil dependency parsing: results using rule based and corpus based approaches', in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 82–95. Berlin, Heidelberg: Springer.

Ramasamy, L. & Žabokrtský, Z. (2012, May) 'Prague dependency style treebank for Tamil', *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1888–1894.

Sarveswaran, K. (2024) 'Building Tamil Treebanks', *arXiv preprint* arXiv:2409.14657.

Sarveswaran, K., Dias, G. & Butt, M. (2021) 'Thamizhi Morph: A morphological parser for the Tamil language', *Machine Translation*, 35(1), pp. 37–70.

Schiffman, H.F. (1999) *A Reference Grammar of Spoken Tamil*. Cambridge University Press.

Sri, J.J. (2023) 'Veṇpāvin tōṟṟamum vaḷarcciyum (in Tamil): Origin and growth of Venpa', *Journal of Tamil Culture and Literature*, 2(4), pp. 37–44.

Subrahmanya Sastri, P.S. (1934) *History of Grammatical Theories in Tamil*. Madras: The Madras Law Journal Press.

Svoboda, E. & Ševčíková, M. (2024, March) 'Compounds in Universal Dependencies: A survey in five European languages', *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 88–99.

Venkatsubhramaniyen, S., Rashmi, S. & Sridhar, R. (2013, February) 'Classification of Tamil poetry using context-free grammar using Tamil grammar rules', *CS & IT Conference Proceedings*, 3(6).